The Design of Sem-Syn Initial Grammar
In Chinese Grammatical Inference

Hsue-Hueh Shih

Department of Foreign Languages and Literature
National Sun Yat-sen University, Taiwan, R.O.C.
E-mail address: hsuehueh@mail.nsysu.edu.tw

Abstract

This paper describes our on-going project on grammatical inference for Chinese. We here emphasize on the design of our sem-syn initial grammar that is a set of stochastic context-free rules and whose probabilistic parameters will be iteratively re-estimated in a corpus-based inference technique. Manually developing and maintaining a grammar for a NLP system has long been regarded as a painful and endless job. Besides, this conventional approach usually results in a grammar with limited coverage. With large bodies of text corpora available on computers, corpus-based grammatical inference (GI) techniques seem to provide a promising solution to the problems. An initial grammar is one of the important components in GI techniques and its function is to facilitate the inference process to proceed. In this paper, we describe the design of our sem-syn initial grammar and how it corresponds to the information given in Sinica Corpus on which our inference system is based. We also give a brief introduction to our Chinese grammatical inference system, showing how the system will use the sem-syn initial grammar to generalize structure from the Corpus.

1. Introduction

Grammar plays an indispensable role in most NLP systems. Conventional handcrafted approach to grammar construction and maintenance has been regarded as painful and endless work. Besides, this conventional approach usually results in a grammar with limited coverage. With the increasing availability of large text corpora in machine-readable form, Grammatical Inference (GI) [Pereira and Schabes, 1992] has surged to provide a promising solution to the problems. In GI, there are two components which are essential to the inference process, namely inference algorithms and initial grammars. An inference algorithm takes the responsibility of learning grammatical knowledge from a set of language samples; whereas an initial grammar is regarded as seed knowledge needed for the inference process to proceed. A well-designed initial grammar can assist the inference algorithm to produce a generalized grammar.

Initial grammars utilized in GI may be classified into four types. First, a null grammar, or an empty grammar, learns all its rules gradually from a set of training sentences in the course of the inference process. Using this approach, inference systems usually start to learn simple grammatical structures from short sentences. This is done by sorting the training data by length and presenting them to the inference systems in an ascending length order [Carroll and Charniak, 1992]. The main disadvantage of this approach is that it tends to acquire a large number of specific rules rather than a set of general ones. This lack of generality leads to an uncontrollable grammar size.

Second, a seed grammar, or core grammar, consists of a set of manually produced rules. Acquisition of new rules proceeds in the course of parsing the training sentences. The pattern of required new rules is often linguistically restricted in order to narrow down the number of plausible candidates [Osborne and Bridge, 1994] [Hindle, 1989]. One of the limitations in this method is that only one parse analysis is allowed to trigger the acquisition process. As a result, new rules are acquired based solely on the first parse available. This conflicts the fact that natural language is ambiguous, and thus all possible analyses should be taken into consideration.

Third, an initial grammar may consist of all possible rules, which are generated using a predefined set of non-terminals and terminals. The form of the rules is usually limited in Chomsky Normal Form and each rule is given a random initial probability. The inference process iteratively modifies the probabilities according to the frequency of use of the rules in the parses of a training set [Pereira and Schabes, 1992]. One disadvantage of this approach is the arbitrary non-terminal labeling of the inferred grammar, which may be linguistically implausible and therefore may weaken the ability for subsequent interpretation of analyzed sentences.

An alternative to these three types of initial grammar is the hybrid grammar used in the Explicit-Implicit technique [Briscoe and Waegner, 1992] [Shih, Young and Waegner, 1995]. The hybrid grammar consists of two sets of production rules, namely explicit and implicit. The explicit rules, like the core grammar mentioned above, are manually produced; the implicit part is similar to the third type of initial grammar but with headedness constraint [Jackendoff, 1977] imposed on the rules. The former aims to analyze general syntactic structure of the target language, whereas the later is responsible for analyzing the sentences the former fails to generate, including the ill-formed. This hybrid initial grammar aims to obtain the merits of the previous two approaches.

In Section 2, we describe the design of our sem-syn initial grammar that is a modified hybrid grammar tailored for Chinese inference. This is followed by in a brief introduction to our Chinese inference system in Section 3, and our conclusion and future work outlined in Section 4.

2. The Design of Sem-Syn Hybrid Grammar

In our system, parts-of-speech rather than words in Sinica Corpus are used in the inference process; therefore, before designing the initial grammar, it is important to examine the POS set utilized in the corpus. There are two issues to be taken into consideration here.

The Chinese Knowledge Information Processing (CKIP) group in Academia Sinica used to classify Chinese words into 178 parts-of-speech for their dictionary of 80,000 entries [CKIP, 1993]. This large set of POSs aims to describe the phenomenon in detail that semantic interpretation of Chinese words has strong influence on the structures of sentences. For instance, “房子” in the sentence “房子蓋好了” is originally the object of the verb “蓋好”, but is moved to the beginning of the sentence because it is a definite noun (we know which house it is). This large set of POS was later reduced to 46 for Sinica Corpus [CKIP, 1995]. This reduction was done by merging some semantically similar words into one. For instance, the semantically intransitive verb “重” in the sentences “這箱子很重” and “這箱子重十公斤” was originally given two different POSs (vh11 and vh12 respectively) because of their different syntactic behaviors. Nevertheless, these two together with other five POSs carrying different syntactic forms of the verb were assigned the same POS, vh, in the corpus.

The second issue is about words carrying different syntactic forms as arguments. In Sinica Corpus, a verb, which takes various syntactic forms as its arguments, is not given different

POSs if the semantics interpretation of the verb in those forms does not change. Here, we call it one-word-one-tag policy. According to this policy, if a verb can take both a clause (the verb is called 句賓述詞) and a noun phrase (the verb is called 單賓述詞) as its arguments, it will only be given one POS which carries a larger element(here it is marked as 句賓述詞). For instance, the verb “討論” can take both a noun phrase and a clause as the arguments in the sentences “我們討論明年的計畫” and “我們討論明年大家是否能申請計畫”, although its POS in the corpus is ve(句賓述詞).

A sem-syn initial grammar in our system is designed to handle the complexities mentioned above. Similar to the Explicit-Implicit technique, the grammar is divided into two components: semantically-oriented and syntactically-oriented rules. The semantically-oriented part of the initial grammar, which is manually developed, is the core part of the grammar responsible for capturing general semantically-consistent structures. For instance, if a verb is classified as active intransitive verb (va: 動作不及物述詞) in the corpus, there will be a semantically-oriented rule: VP -->va, regardless of its syntactic behavior or its other possible POSs. The following is an example showing how the rules look like:

```
VP1 [active +] --> va       /* 大軍出動了    */
VP1 [active +] --> vb PP    /* 授粉給雌蕊    */
VP1 [active -] --> vh       /* 這箱子很重    */
VP1 [active -] --> vi PP    /* 醉心於政治    */
```

The feature *active* is utilized to indicate whether the verb is an active (動作) or a stative (狀態) verb. Note that the stative intranstive verb vh also has other syntactic form that carries an NP (十公斤) as mentioned earlier in this section, but we leave it to the syntactically-oriented rules to handle since its form is inconsistent with the definition of vh.

The syntactically-oriented part, designed to generate from rule templates, is to handle the syntactic behaviors of a verb, accommodate the structures that are excluded due to the one-word-one-tag policy, and even deal with ill-formed sentences in the corpus. Like implicit rules with headedness constraint in the Explicit-Implicit technique, the syntactically-oriented rules will be generated from the following templates:

```
NT --> NT NT
NT --> NT T
NT --> T NT
NT --> T T
```

Where NT is a non-terminal and T is a terminal (part-of-speech) symbols used in the grammar. However, implicit rules are a set of non-recursive rules with limited generating power. It is believed that the syntactically-oriented rules need to handle the majority of the training data due to the complexity of POSs in the corpus mentioned at the beginning of this section. Therefore, it is desirable to loosen the headedness constraint so that the bar level of a head daughter can be equal to or less than that of its mother. This results in a set of recursively syntactically-oriented rules.

The two set of rules will be put together to form our sem-syn initial grammar with a set of corresponding random probabilities. This grammar will then be ready for our inference process to proceed.

3. The Chinese Grammatical Inference System

We are currently developing a corpus-based grammatical inference system for Chinese.

The system consists of three components:

- Initial Grammar

The sem-syn hybrid grammar mentioned in the previous section is currently under development. A grammar development environment tool called GDE [Carroll, Briscoe and Grover 1991] is employed to develop our semantically-oriented rules in GPSG formalism, whereas syntactically-oriented rules will be generated using four templates shown in Section 2. Both parts of rules will then be converted into context-free rules and given random initial probabilities to meet the requirement of the stochastic inference algorithm mentioned below.

- Corpus

The pre-tagged primary school textbooks from Sinica Corpus are used for both training and testing. These data will be manually phrase-bracketed as a tree bank to provide the phrasal information during inference process. The phrase-bracketed test set will be used to examine the bracketing accuracy of the parses generated by the inferred grammar.

- Inference Algorithm

The system utilizes a chart-based Inside-Outside algorithm [Waegner, 93]. It is a stochastic inference algorithm that can take a stochastic context-free grammar as a source and iteratively re-estimates the set of probabilistic parameters of the grammar. Analogous to the forward and backward probabilities in conventional Hidden Markov Model(HMM), this algorithm define the inside(e) and outside(f) probabilities as:

$$e(s,t,I) = P(S \overset{*}{=}> O(s),\cdots..,O(t)/G),$$

$$f(s,t,I) = P(S \overset{*}{=}> O(1),\cdots,O(s-1),I,O(t+1),\cdots,O(T)/G)$$

where e(s,t,I) is the probability of the non-terminal symbol I generating the observation O(s),…·.,O(t), and f(s,t,I) the probability of I being generated but not involved in generating the observation O(1),…,O(s-1), and O(t+1),…,O(T). G is the grammar, T is the total number of elements in the observation O(1),…·.,O(T), and $1 \leqq s \leqq t \leqq T$. The inside probabilities are computed bottom-up, and outside probabilities are computed top-down. Like training the transition and emission probabilities in HMM, the values of e and f of non-terminal symbols can be used to re-estimate rule probabilities of the grammar in the similar fashion.

In our system, the set of probabilistic parameters of the sem-syn grammar will be iteratively re-estimated from all legitimate parses that conform with the bracketing constraints in our training tree bank (it is called supervised training). The inference process finishes when a change in total log probability of training sentences is less than a set threshold.

4. Conclusion and Future Work

We have outlined our on-going research on the design of the sem-syn initial grammar for the Chinese inference system. Unlike its European counterparts, Chinese language has its structure complexity, which have lead to a different design on the initial grammar for grammatical inference.

The sem-syn initial grammar has been under development in the project. We will soon start to build up the tree bank for the supervised training and develop the inference system. We hope that the inferred grammar will not only reflect the sentence structure in the training set, but also can predict the unseen sentences (test data) which in some sense are of the same nature as the training data. This expectation will be verified by experiments in which the test sentences are analyzed using the inferred grammar and their results (parses) are examined using the test tree bank.

References

Briscoe, E. and Waegner, N. (1992) Robust stochastic parsing using Inside-Outside algorithm. In AAAI Symposium on Statistic Applications to Natural Language.

Carroll, G. and Charniak; E. (1992) Learning probabilistic dependency grammars from labelled text. In AAAI fall Symposium Series: Probabilistic Approach to Natural Language, pp. 25-32.

Carroll, J., Briscoe, E., and Grover, C. (1991) A development environment for large natural language grammars. TR. 233, Computer Laboratory, Cambridge University, England.

CKIP (Chinese Knowledge Information Processing Group), (1995) Introduction to Sinica Corpus: A tagged balance corpus for Mandarin Chinese.TR95-02. Taipei: Academia Sinica.

CKIP (Chinese Knowledge Information Processing Group), (1993) Analysis of Chinese parts-of-speech.TR93-05. Taipei: Academia Sinica.

Hindle, D. (1989) Acquiring disambiguation rules from text. In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, pp. 118-125.

Osborne, M. and Bridge, D. (1994) Learning unification-based grammars using the spoken English corpus. In R.C. Carrosco and J. Oncina, editors, Proceedings of the Second International Colloquium on Grammatical Inference, pp. 260-270, Alicante. Springer-Verlag.

Pereira F. and Schabes Y, (1992). Inside-Outside re-estimation for partially bracketed corpora. In Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, pp. 128-135.

Shih, H-H., Young, S., and Waegner, N. (1995) An inference approach to grammar construction. Computer Speech and Language, 9:235-256

Waegner, N. (1993). Stochastic Model for Language Acquisition. PhD thesis, Cambridge University, England.